International Journal of Multimedia and Ubiquitous Engineering
Vol.8, No.6 (2013), pp.105-118
http://dx.doi.org/10.14257/ijmue.2013.8.6.11

RMF: Rough Set Membership Function-based for Clustering Web Transactions

Tutut Herawan[1] and Wan Maseri Wan Mohd[2]

[1]*Department of Mathematics Education, Universitas Ahmad Dahlan*
Jalan Prof Dr Soepomo 55166, Yogyakarta, Indonesia
[2]*Faculty of Computer System and Software Engineering*
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Pahang, Malaysia

tutut81@uad.ac.id, maseri@ump.edu.my

Abstract

One of the most important techniques to improve information management on the web in order to obtain better understanding of user's behaviour is clustering web data. Currently, the rough approximation-based clustering technique has been used to group web transactions into clusters. It is based on the similarity of upper approximations of transactions to merge between two or more clusters. However, in reviewing the technique, it has a weakness in terms of processing time in obtaining web clusters. In this paper, an alternative technique for grouping web transactions using rough set theory, named RMF is proposed. It is based on the rough membership function of a transaction similarity class with respect to the other classes. The two UCI benchmarks datasets are opted in the experimental processes. The experimental results reveal that the proposed technique has an benefit of low time complexity as compared to the baseline technique up to 67 %.

Keywords: *Clustering; Web transactions; Rough membership function; Rough set theory*

1. Introduction

As one of the most important tasks of Web Usage Mining (WUM), web user clustering, which establishes groups of users exhibiting similar browsing patterns, provides useful knowledge to personalized web services and motivates long term research interests in the web community (Ling, *et al.*, 2009). Existing web usage data mining techniques include statistical analysis (Srivastava, *et al.*, 2000), association rules (Huang, *et al.*, 2002; Mobasher, *et al.*, 2001), sequential patterns (Yang, *et al.*, 2001), classification (Li, *et al*, 2001), and clustering (Labroche, *et al*, 2001; Mobasher, *et al.*, 1999).The goal of clustering is to create groups of data objects in an unsupervised fashion so that data items in the same cluster are similar to each other, yet dissimilar to data items residing in other clusters (Morteza, *et al.*, 2008). Generally, web users may exhibit various types of behaviours associated with their information needs and intended tasks when they are traversing the Web. These task-oriented behaviours are explicitly characterized by sequences of clicks on different web items performed by users. As a result, these tasks are implicitly captured by inducing the underlying relationships among the click stream data.

Access transaction over the web can be expressed in the two finite sets, user transaction and hyperlinks/URLs (De & Krishna, 2004). A user transaction U is a sequence of items, this set is formed by m users and the set A is a set of distinct n clicks (hyperlinks/URLs) clicked

ISSN: 1975-0080 IJMUE

by users that are $U = \{t_1, t_2, \cdots, t_m\}$ and $A = \{hl_1, hl_2, \cdots, hl_n\}$ where for every $t_i \in T \subseteq U$ is a non-empty subset of *U*. The temporal order of user clicks within transactions has been taken into account. A user transaction $t \in T$ is represented as a vector $t = \{u_1^t, u_2^t, \cdots, u_n^t\}$, where $u_i^t = 1$ if $hl_i \in t$ and $u_i^t = 0$ if otherwise.

A well-known approach for clustering web transactions is using rough set theory (Pawlak, 1982; Pawlak, 1991; Pawlak & Skowron, 2007). De and Krishna (De & Krishna, 2004) proposed an algorithm for clustering web transactions using rough approximation. It is based on the similarity of upper approximations of transactions by given any threshold. However, there are some iterations should be done to merges of two or more clusters that have the same similarity of upper approximations. To overcome the problem, in this paper, we propose an alternative technique for clustering web transaction. It is based on the rough membership function of a transaction similarity class with respect to the other classes. In summary, there are three contributions of this work:

a. We use the concept of the rough membership function. The function takes on a transaction similarity class with respect to the other classes.

b. We show that the proposed technique differs on how to allocate transaction in the same cluster.

c. We show that the proposed technique produce lower time complexity as compared with that (De & Krishna, 2004).

The rest of the paper is organized as follows. Section 2 describes the concept of rough set theory. Section 3 describes analysis of the baseline technique on web transaction clustering proposed by (De & Krishna, 2004). Section 4 describes the proposed technique. Section 5 describes the experimental tests on two UCI benchmark datasets. Finally, we conclude our works in Section 6.

2. Rough Set Theory

In the 1980's, Pawlak introduced rough set theory to deal the problem of vagueness and uncertainty in datasets (Pawlak, 1982). Similarly to fuzzy set theory, it is not an alternative to classical set theory but it is embedded in it. Fuzzy and rough sets theories are not competitive, but complementary to each other (Pawlak & Skowron, 2007). Rough set theory has attracted attention to many researchers and practitioners all over the world, who contributed essentially to its development and applications (Herawan & Deris, 2009a; Herawan & Deris, 2009c; Herawan & Deris, 2009d; Herawan, *et al.*, 2009; Herawan, *et al.*, 2011a; Herawan, *et al.*, 2011b; Herawan, *et al.*, 2011c; Herawan, *et al*, 2011d; Senan, *et al.*, 2011a; Senan, *et al.*, 2011b; Yanto, *et al.*, 2010a; Yanto, *et al.*, 2010b; Yanto, *et al.*, 2011; Yanto, *et al*, 2012). The original goal of the rough set theory is induction of approximations of concepts. The idea consists of approximation of a subset by a pair of two precise concepts called the *lower approximation* and *upper approximation*. Intuitively, the lower approximation of a set consists of all elements that surely belong to the set, whereas the upper approximation of the set constitutes of all elements that possibly belong to the set. The difference of the upper approximation and the lower approximation is a *boundary region*. It consists of all elements that cannot be classified uniquely to the set or its complement, by employing available knowledge. Thus any rough set, in contrast to a crisp set, has a non-empty boundary region. Motivation for rough set theory has come from the need to represent a subset of a universe in terms of equivalence classes of a partition of the universe. In this chapter, the basic concept of rough set theory in terms of data is presented.

2.1. Information System

Data are often presented as a table, columns of which are labelled by *attributes*, rows by *objects* of interest and entries of the table are *attribute values*. By an *information system*, we mean a 4-tuple (quadruple) $S=(U,A,V,f)$, where U is a non-empty finite set of objects, $A$is a non-empty finite set of attributes, $V=\bigcup_{a\in A}V_a$, V_a is the domain (value set) of attribute *a*, $f:U\times A\to V$ is a total function such that $f(u,a)\in V_a$, for every $(u,a)\in U\times A$, called information (knowledge) function (Herawan & Deris, 2009b). An information system is also called a knowledge representation systems or an attribute-valued system and can be intuitively expressed in terms of an information table (refer to Table 1).

Table 1. An information system

U/A	a_1	a_2	⋯	a_k	⋯	$a_{\lvert A\rvert}$
u_1	$f(u_1,a_1)$	$f(u_1,a_2)$	⋯	$f(u_1,a_k)$	⋯	$f(u_1,a_{\lvert A\rvert})$
u_2	$f(u_2,a_1)$	$f(u_2,a_2)$	⋯	$f(u_2,a_k)$	⋯	$f(u_2,a_{\lvert A\rvert})$
u_3	$f(u_3,a_1)$	$f(u_3,a_2)$	⋯	$f(u_3,a_k)$	⋯	$f(u_3,a_{\lvert A\rvert})$
⋮	⋮	⋮	⋱	⋮	⋱	⋮
$u_{\lvert U\rvert}$	$f(u_{\lvert U\rvert},a_1)$	$f(u_{\lvert U\rvert},a_2)$	⋯	$f(u_{\lvert U\rvert},a_k)$	⋯	$f(u_{\lvert U\rvert},a_{\lvert A\rvert})$

2.2. Indiscernibility relation

The starting point of rough set theory is the indiscernibility relation, which is generated by information about objects of interest. The indiscernibility relation is intended to express the fact that due to the lack of knowledge we are unable to discern some objects employing the available information. Therefore, generally, we are unable to deal with single object. Nevertheless, we have to consider clusters of indiscernible objects. The following definition precisely defines the notion of indiscernibility relation between two objects.

Definition 2.1. *Let $S=(U,A,V,f)$ be an information system and let B be any subset of A. Two elements $x,y\in U$ are said to be B-indiscernible (indiscernible by the set of attribute $B\subseteq A$ in S) if and only if $f(x,a)=f(y,a)$, for every $a\in B$.*

Obviously, every subset of *A* induces unique indiscernibility relation. Notice that, an indiscernibility relation induced by the set of attribute *B*, denoted by $IND(B)$, is an equivalence relation. It is well known that, an equivalence relation induces unique partition. The partition of U induced by $IND(B)$ in $S=(U,A,V,f)$ denoted by U/B and the equivalence class in the partition U/B containing $x\in U$, denoted by $[x]_B$.

Given arbitrary subset $X\subseteq U$, *X* may not be presented as union of some equivalence classes in *U*. In other means that a subset *X* cannot be described precisely in $S=(U,A,V,f)$. Thus, a subset *X* may be characterized by a pair of its approximations, called lower and upper approximations. It is here that the notion of rough set emerges.

2.3. Set Approximations

The indiscernibility relation will be used to define set approximations that are the basic concepts of rough set theory. The notions of lower and upper approximations of a set can be defined as follows.

Definition 2.2. *Let* $S=(U,A,V,f)$ *be an information system, let B be any subset of A and let X be any subset of U. The B-lower approximation of X, denoted by* $\underline{B}(X)$ *and B-upper approximations of X, denoted by* $\overline{B}(X)$*, respectively, are defined by*

$$\underline{B}(X)=\{x\in U \mid [x]_B \subseteq X\} \text{ and } \overline{B}(X)=\{x\in U \mid [x]_B \cap X \neq \phi\}.$$

The accuracy of approximation (accuracy of roughness) of any subset $X\subseteq U$ with respect to $B\subseteq A$, denoted $\alpha_B(X)$ is measured by

$$\alpha_B(X)=\frac{|\underline{B}(X)|}{|\overline{B}(X)|}, \tag{3.1}$$

where $|X|$ denotes the cardinality of *X*. For empty set ϕ, it is defined that $\alpha_B(\phi)=1$ (Pawlak & Skowron, 2007). Obviously, $0\le\alpha_B(X)\le 1$. If *X* is a union of some equivalence classes of *U*, then $\alpha_B(X)=1$. Thus, the set *X* is *crisp* (precise) with respect to *B*. And, if *X* is not a union of some equivalence classes of *U*, then $\alpha_B(X)<1$. Thus, the set *X* is *rough* (imprecise) with respect to *B* (Pawlak & Skowron, 2007). This means that the higher of accuracy of approximation of any subset $X\subseteq U$ is the more precise (the less imprecise) of itself.

2.4. Rough membership function

Rough sets can be also defined employing, instead of approximation, rough membership function [12, 13] as follow

$\mu_X^B : U\to[0,1]$, where $\mu_X^B(x)=\frac{|X\cap[x]_B|}{|[x]_B|}$ and $|X|$ denotes the cardinality of *X*.

The rough membership function expresses conditional probability that *x* belongs to *X* given *R* and can be interpreted as a degree that *x* belongs to *X* in view of information about *x* expressed by *B*. The meaning of rough membership function can be depicted as shown in Figure 1.

The rough membership function can be used to define approximations and the boundary region of a set, as shown below

$$\underline{B}(X)=\{x\in U : \mu_X^B(x)=1\} \text{ and } \overline{B}(X)=\{x\in U : \mu_X^B(x)>0\}.$$

Example 2.2. Consider the following information system as in Table 2. Suppose we are given data about 6 students, as shown below.

Table 2. A decision system

Student	Analysis	Algebra	Statistics	Decision
1	bad	good	medium	accept
2	good	bad	medium	accept
3	good	good	good	accept
4	bad	good	bad	reject
5	good	bad	medium	reject
6	bad	good	good	accept

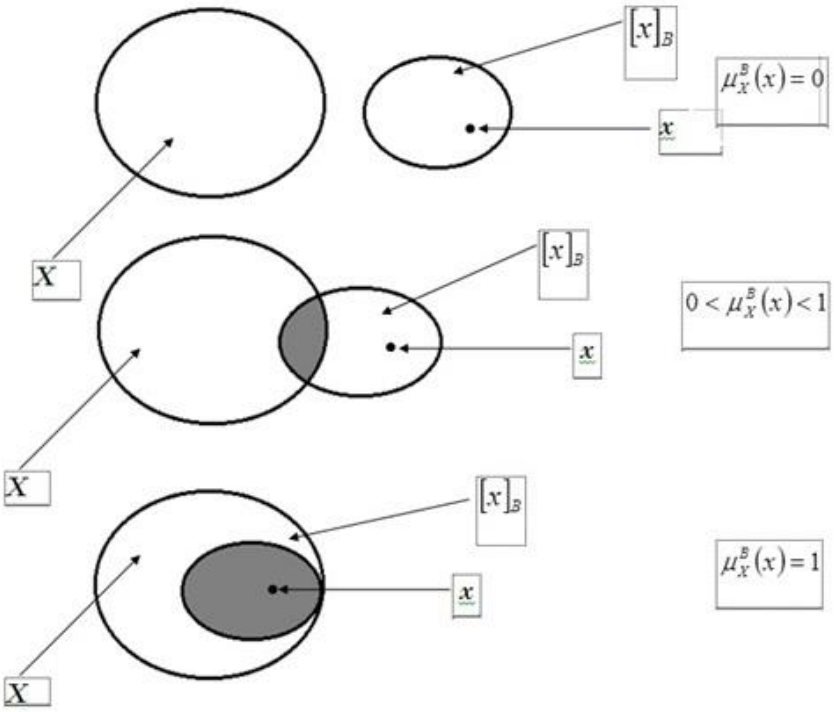

Figure 1. Rough membership functions

From Table 2, we have

$$U=\{1,2,3,4,5,6\},$$
$$A=\{\text{Analysis, Algebra, Statistics}\}=C\bigcup\{\text{Decision}\}=D,$$
$$V_{\text{Analysis}}=\{\text{bad, good}\},$$
$$V_{\text{Algebra}}=\{\text{bad, good}\},$$
$$V_{\text{Statistics}}=\{\text{bad, medium, good}\},$$
$$V_{\text{Decision}}=\{\text{accept,reject}\}.$$

For the set of condition attributes, $C=\{\text{Analysis, Algebra, Statistics}\}$, we have

$$U/C=\{\{1\},\{2,5\},\{3\},\{4\},\{6\}\}.$$

Let $X(\text{Decision}:\text{accept})=\{1,2,3,6\}$. Thus,

$$\text{For } x=1, \text{ then } \mu_X^C(x)=\frac{|[x]_C \bigcap X|}{|[x]_C|}=\mu_{\{1,2,3,6\}}^C(1)=\frac{|\{1\}\bigcap\{1,2,3,6\}|}{|\{1\}|}=1,$$

$$\text{For } x=2, \text{ then } \mu_X^C(x)=\frac{|[x]_C \bigcap X|}{|[x]_C|}=\mu_{\{1,2,3,6\}}^C(2)=\frac{|\{2,5\}\bigcap\{1,2,3,6\}|}{|\{2,5\}|}=0.5,$$

$$\text{For } x=3, \text{ then } \mu_X^C(3)=\frac{|[x]_C \bigcap X|}{|[x]_C|}=\mu_{\{1,2,3,6\}}^C(3)=\frac{|\{3\}\bigcap\{1,2,3,6\}|}{|\{3\}|}=1,$$

$$\text{For } x=6, \text{ then } \mu_X^C(6)=\frac{|[x]_C \bigcap X|}{|[x]_C|}=\mu_{\{1,2,3,6\}}^C(6)=\frac{|\{6\}\bigcap\{1,2,3,6\}|}{|\{6\}|}=1.$$

3. Analysis of Data Clustering Technique Proposed by (De & Krishna, 2004)

In this section, we discuss the technique proposed by (De & Krishna, 2004). Given two transactions *s* and *t*, the measurement of similarity between *t* and *s* is given by

$$\text{sim}(s,t)=\frac{|s\cap t|}{|s\cup t|}$$

Obviously, $\text{sim}(s,t)\in[0,1]$, where $\text{sim}(s,t)=1$, when two transactions *s* and *t* are exactly identical and $\text{sim}(s,t)=0$, when two transactions *s* and *t* have no items in common. De and Krishna (De & Krishna, 2004) used a binary relation *R* on *T* defined as follows.

For any threshold value $\text{th}\in[0,1]$ and for any two user transactions $s,t\in T$, a binary relation *R* on *T* denoted as *sRt* if and only if $\text{sim}(s,t)\geq \text{th}$. This relation *R* is a tolerance relation as *R* is both reflexive and symmetric, but transitive may not hold good always.

Definition 3.1. *The similarity class of t, denoted by R(t), is a set of transactions which are similar to t which is given by* $R(t)=\{s\in T: sRt\}$.

For different threshold values, one can get different similarity classes. A domain expert can choose the threshold based on this experience to get a proper similarity class. It is clear that for a fixed threshold $\in[0,1]$, a transaction form a given similarity class may be similar to an object of another similarity class.

Definition 3.2. *Let* $P\subseteq T$*, for a fixed threshold* $\text{th}\in[0,1]$ *a binary tolerance relation R is defined on T. The lower approximation of P, denoted by* $\underline{R}(P)$ *and the upper approximation of P, denoted by* $\overline{R}(P)$ *are respectively defined as*

$$\underline{R}(P)=\{t\in P: R(t)\subseteq P\} \text{ and } \overline{R}(P)=\bigcup_{t\in P}R(t).$$

De and Krishna (De & Krishna, 2004) proposed a technique of clustering the clicks of user navigations called as similarity upper approximation has been and denoted by S_i. A set of transactions that are possibly similar to $\overline{R}(t_i)$ is denoted by $\overline{RR}(t_i)$. This process continues until two consecutive upper approximations for t_i; for $i = 1,2,\cdots,|U|$ are the same and two or more clusters that have the same similarity upper approximations merges at each iteration.

With this technique, we need high computational complexity to cluster the transactions. This is due to find out the similarity upper approximation until two consecutive upper approximations are same. To overcome this problem, we propose an alternative technique to cluster the transactions.

4. The Proposed Technique

The proposed technique for clustering the transactions is used the rough membership function. It is based on membership both two similarity classes.

Definition 4.1. *Let $R(t_i)$ and $R(t_j)$ be similarity classes of the transaction t_i. The rough membership function of similarity classes $R(t_i)$ with respect to $R(t_j)$, where $i \neq j$, denoted by $\mu_{R(t_i)}^{R(t_j)}(t_i)$, where*

$$\mu_{R(t_i)}^{R(t_j)}(t_i): R(t_i) \to [0,1],$$

is defined by

$$\mu_{R(t_i)}^{R(t_j)}(t_i) = \frac{R(t_i) \cap R(t_j)}{R(t_i)}.$$

Where $\mu_{R(t_i)}^{R(t_j)}(t_i) = 0$ if $R(t_i)$ and $R(t_j)$ have void intersection, otherwise $\mu_{R(t_i)}^{R(t_j)}(t_i) \in (0,1]$, especially if $R(t_j) \subseteq R(t_i)$, then $\mu_{R(t_i)}^{R(t_j)}(t_i) = 1$. The membership function $\mu_{R(t_i)}^{R(t_j)}(t_i)$ is a tolerance function as $\mu_{R(t_i)}^{R(t_j)}(t_i) \in (0,1]$ is both reflexive and symmetric.

Definition 4.2. *Two clusters S_i and S_j, for $i \neq j$ are said to be the same if and only if*

$$S_i = \bigcup R(t_i), \text{ for } i = 1,2,\cdots,|U|.$$

Proposition 4.1. *Let S_i and S_j be two clusters. If $\mu_{R(t_i)}^{R(t_j)}(t_i) \in (0,1]$, then $S_i = S_j$.*

Proof. Suppose $S_i \neq S_j$, for $i \neq j$, then we have $S_i \cap S_j = \phi$. From Definition 4.2, then

$$(\cup R(t_i)) \bigcap (\cup R(t_j)) = \phi$$
$$R(t_i) \bigcap R(t_j) = \phi$$

Thus, based on Definition 4.1, if $R(t_i) \bigcap R(t_j) = \phi$, then $\mu_{R(t_i)}^{R(t_j)}(t_i) = 0$. This is contradiction from hypothesis.

Corollary 4.1. *If* $\mu_{R(t_i)}^{R(t_j)}(t_i) \in (0,1]$, *then* $\bigcup R(t_i) = S$.

Proof. It is clear from Proposition 4.1.

4.1. The computational complexity

Suppose that there are n objects in an information system $S = (U, A, V, f)$ of web user transaction. Therefore, there are at most n similarity classes. The technique needs n^2 computation for determining the rough membership function in order to obtain the cluster. Thus, the computational complexity is the polynomial $O(n^2 + n)$.

4.2. Example

In this study, the comparisons between the proposed technique and the technique proposed by (De & Krishna, 2004) are presented by given example. In this example, the data set in (De & Krishna, 2004) is used. De and Khrisna provide a web user transaction containing four objects $|U| = 4$ with five hyperlinks. Let $U = \{t_1, t_2, t_3, t_4\}$ be the set of user transactions and $A = \{hl_1, hl_2, hl_3, hl_4, hl_5\}$ be a set of distinct URLs accessed from the user transaction *U*. Let $t_1 = \{hl_1, hl_2\}$, $t_2 = \{hl_2, hl_3, hl_4\}$, $t_3 = \{hl_1, hl_3, hl_5\}$, and $t_4 = \{hl_2, hl_3, hl_5\}$. Thus, they can be represented in a Boolean-valued information system as in Table 3 below.

Table 3. Data transactions

U / A	hl_1	hl_2	hl_3	hl_4	hl_5
t_1	1	1	0	0	0
t_2	0	1	1	1	0
t_3	1	0	1	0	1
t_4	0	1	1	0	1

The first step of the technique is obtaining the measure of similarity that gives information about the users access patterns related to their common areas of interest by similarity relation between two transactions of objects. The calculation of the measure of similarity user transactions are shown as follow

$$\text{sim}(t_1, t_2) = \frac{|t_1 \cap t_2|}{|t_1 \cup t_2|} = \frac{|\{hl_2\}|}{|\{hl_1, hl_2, hl_3, hl_4\}|} = \frac{1}{4} = 0.25,$$

$$\text{sim}(t_1, t_3) = \frac{|t_1 \cap t_3|}{|t_1 \cup t_3|} = \frac{|\{hl_1\}|}{|\{hl_1, hl_2, hl_3, hl_5\}|} = \frac{1}{4} = 0.25,$$

$$\text{sim}(t_1, t_4) = \frac{|t_1 \cap t_4|}{|t_1 \cup t_4|} = \frac{|\{hl_2\}|}{|\{hl_1, hl_2, hl_3, hl_5\}|} = \frac{1}{4} = 0.25,$$

$$\text{sim}(t_2, t_3) = \frac{|t_2 \cap t_3|}{|t_2 \cup t_3|} = \frac{|\{hl_3\}|}{|\{hl_1, hl_2, hl_3, hl_5\}|} = \frac{1}{4} = 0.25,$$

$$\text{sim}(t_2, t_4) = \frac{|t_2 \cap t_4|}{|t_2 \cup t_4|} = \frac{|\{hl_2, hl_3\}|}{|\{hl_2, hl_3, hl_4, hl_5\}|} = \frac{2}{4} = 0.5,$$
$$\text{sim}(t_3, t_4) = \frac{|t_3 \cap t_4|}{|t_3 \cup t_4|} = \frac{|\{hl_3, hl_5\}|}{|\{hl_1, hl_2, hl_3, hl_5\}|} = \frac{2}{4} = 0.5.$$

The similarity classes can be obtained by given the threshold value using Definition 3.2. In this case, by set the threshold to 0.5. Then, the similarity classes are obtained as in Figure 2.

$$R(t_1) = \{t_1\},$$
$$R(t_2) = \{t_2, t_4\},$$
$$R(t_3) = \{t_3, t_4\},$$
$$R(t_4) = \{t_2, t_3, t_4\}.$$

Figure 2. The similarity classes

The last step is to cluster the transactions. To get the cluster (De & Krishna, 2004), Similarity Upper Approximations is used. For instance, the processes to obtain upper approximation for $\overline{R}(t_2)$ and $\overline{R}\overline{R}(t_2)$ are shown below.

1. To obtain $\overline{R}(t_2)$; since $\overline{R}(P)$, for $P = \{t_2\}$; then

$$t_1 \notin P \text{ for class } R(t_1),$$
$$t_2 \in P \text{ for class } R(t_2),$$
$$t_3 \notin P \text{ for class } R(t_3),$$
$$t_4 \notin P \text{ for class } R(t_4),$$
$$\overline{R}(t_2) = \bigcup R(t_2) = \{t_2, t_4\}.$$

2. To obtain $\overline{R}\overline{R}(t_2)$; since $\overline{R}\overline{R}(P)$, for $P = \{t_2\}$; then

$$t_1 \notin P \text{ for class } \overline{R}(t_1),$$
$$t_2 \in P \text{ for class } \overline{R}(t_2),$$
$$t_3 \notin P \text{ for class } \overline{R}(t_3),$$
$$t_4 \notin P \text{ for class } \overline{R}(t_4),$$
$$\overline{R}\overline{R}(t_2) = \bigcup \{R(t_2), R(t_4)\} = \{t_2, t_3, t_4\}.$$

The overall similarity upper approximation processes are shown as follows:
First iteration

$$\overline{R}(t_1) = \{t_1\},$$
$$\overline{R}(t_2) = \{t_2, t_4\},$$

$$\overline{R}(t_3) = \{t_2, t_4\},$$
$$\overline{R}(t_4) = \{t_2, t_3, t_4\}.$$

Second iteration

$$\overline{R}\overline{R}(t_1) = \{t_1\},$$
$$\overline{R}\overline{R}(t_2) = \{t_2, t_3, t_4\},$$
$$\overline{R}\overline{R}(t_3) = \{t_2, t_3, t_4\},$$
$$\overline{R}\overline{R}(t_4) = \{t_2, t_3, t_4\}.$$

Third iteration

$$\overline{R}\overline{R}\overline{R}(t_1) = \{t_1\},$$
$$\overline{R}\overline{R}\overline{R}(t_2) = \{t_2, t_3, t_4\},$$
$$\overline{R}\overline{R}\overline{R}(t_3) = \{t_2, t_3, t_4\},$$
$$\overline{R}\overline{R}\overline{R}(t_4) = \{t_2, t_3, t_4\}.$$

Based on all iterations shown previously, it is observed that two consecutive upper approximation for t_1, t_2 , t_3, and t_4 are similar. Thus, the similarity upper approximation for t_1, t_2 , t_3, and t_4 are shown below

$$S_1 = \{t_1\},\ S_2 = \{t_2, t_3, t_4\},\ S_3 = \{t_2, t_3, t_4\},\ \text{and}\ S_4 = \{t_2, t_3, t_4\}.$$

where $S_2 = S_3 = S_4$ and $S_1 = S_i$, for $i = 2,3,4$ and finally resulted in these two clusters

$$\{t_1\} \text{ and } \{t_2, t_3, t_4\}.$$

Obviously, some iteration should be done in order to find the similarity of upper approximations of a transaction and then merge of two or more clusters that have the same similarity of upper approximations.

However, the proposed technique uses the rough membership function. The rough membership function of similarity classes $R(t_2)$ with respect to $R(t_1)$, $R(t_3)$ and $R(t_4)$ are shown as follow

$$\mu_{R(t_2)}^{R(t_1)}(t_2) = \frac{R(t_1) \cap R(t_2)}{R(t_2)} = \frac{|\{t_1\} \cap \{t_2, t_4\}|}{|\{t_2, t_4\}|} = \frac{|\phi|}{|\{t_2, t_4\}|} = \frac{0}{2} = 0,$$
$$\mu_{R(t_2)}^{R(t_3)}(t_2) = \frac{R(t_3) \cap R(t_2)}{R(t_2)} = \frac{|\{t_3, t_4\} \cap \{t_2, t_4\}|}{|\{t_2, t_4\}|} = \frac{|\{t_4\}|}{|\{t_2, t_4\}|} = \frac{1}{2} = 0.5,$$
$$\mu_{R(t_2)}^{R(t_4)}(t_2) = \frac{R(t_4) \cap R(t_2)}{R(t_2)} = \frac{|\{t_2, t_3, t_4\} \cap \{t_2, t_4\}|}{|\{t_2, t_4\}|} = \frac{|\{t_4\}|}{|\{t_2, t_4\}|} = \frac{2}{2} = 1.$$

The similar calculations are performed for all the transactions. These calculations are shown in Table 3.

Table 4. Rough membership value of similarity classes

$R(t)$	1	2	3	4
1	-	0	0	0
2	0	-	0.5	1
3	0	0.5	-	1
4	0	1	1	-

Here, since $\mu_{R(t_i)}^{R(t_j)}(t_i) \in (0,1]$ for $i \neq j$ and $i, j = 2,3,4$, the clusters are obtained as

$$S_1 = \bigcup R(t_1) = \{t_1\} \quad S_i = \bigcup R(t_i) = \{t_2, t_3, t_4\} \text{ for } i = 2,3,4.$$

Hence, the two clusters are $\{t_1\}$ and $\{t_2, t_3, t_4\}$. These are the same clusters which that in (De & Krishna, 2004). However, the iteration is lower than that of the technique proposed by (De & Krishna, 2004). Therefore, the proposed technique to clusters the transactions perform better than that (De & Krishna, 2004).

5. Results and Discussion

In order to test the proposed technique and compare it with the technique of (De & Krishna, 2004), the two UCI benchmark datasets taken from: http:/kdd.ics.uci.edu/databases/msnbc/msnbc.html and http:/kdd.ics.uci.edu/databases/Microsoft/microsoft.html are opted in the simulation processes. Msnbc datasets data set the page visited by user on September 28, 1999. Visitors are recorded at the level of URL category chronologically. The data is taken from mcnbc.com Internet Information Server (IIS) logs. Each row of the data set corresponds to a request of a user for a page. While, Microsoft datasets describes the page visited (*www.microsoft.com*) by user on February, 1998. The data lists all the pages of the web site (Vroots) that each of the user visited in a week timeframe. Vroots are identified by their title and URL. The client-side cached data is not recorded, thus this data contains only the server-side log. From almost one million transactions, only first 2000 transactions are used and then they are split into five categories, namely; 100, 200, 500, 1000, and 2000 transactions.

The proposed technique for clustering web transactions is implemented in MATLAB version 7.10.0.499 (R2010a). They are executed sequentially on a processor AMD Turion 64×2, MMX, 3DNow, (2CPUs).The total main memory is 1 Gigabyte and the operating system is Windows XP Professional SP2.

Table 5 and Table 6 show the performance comparison in term of processing time in second and computational. Based on the result, the proposed approach shows improvement of more than 60 % for transactions ranging from 100 to 2000. The approach by (De & Krishna, 2004) took longer time since it discovers the similarity of upper approximations. The Proposed approach taken less processing time since only the membership of two similarity classes are used.

Table 5. Executing time on Msnbc dataset

Number of Transactions	RMF	D & K (De & Krishna, 2004)	Improvement
100	2.578	6.703	61.54 %
200	10.735	31.672	66.11 %
500	104.797	326.110	67.86 %
1000	838.376	2608.880	67.86 %
2000	2934.316	9131.080	67.86 %
Average			66.25 %

Table 6. Executing time on Microsoft dataset

Number of Transactions	RMF	D & K (De & Krishna, 2004)	Improvement
100	3.297	9.516	65.35 %
200	15.265	46.984	67.51 %
500	137.385	422.856	67.51 %
1000	1099.08	3382.848	67.51 %
2000	3846.78	11839.968	67.51 %
Average			67.08 %

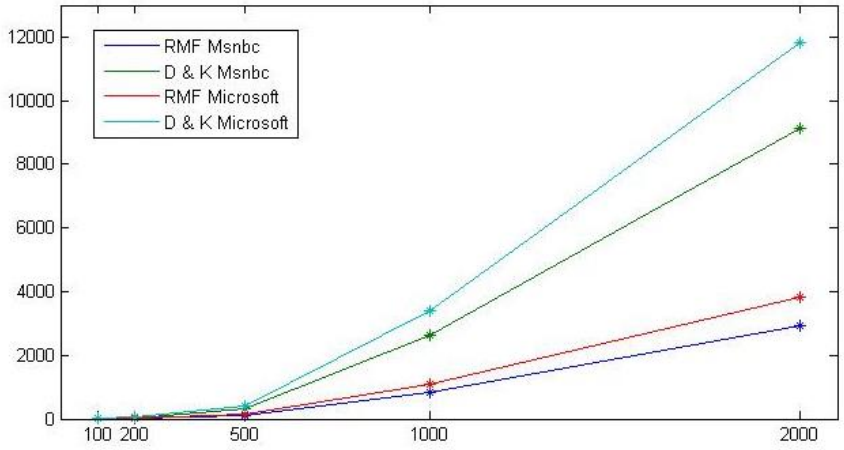

Figure 3. Performance comparison by executing time

6. Conclusion

A web clustering technique can be applied to find interesting user access patterns in web log. In this paper, we have proposed an alternative technique for clustering web transactions using rough membership of similarity class between two transactions. The performance of the proposed technique was presented in terms of processing time. Two sets of benchmark data with 2000 transactions taken from web server through *http://kdd.ics.uci.edu* are used in the simulation processes. It is shown that the proposed technique requires significantly lower response time up to 66.25 % and 67.08 % as compared to the technique of (De & Krishna, 2004), respectively.

References

[1] S. K. De and P. R. Krishna, "Clustering web transactions using rough approximation", Fuzzy Sets and Systems, vol. 148, (**2004**), pp. 131- 138.
[2] T. Herawan and M. M. Deris, "A framework on rough set-based partitioning attribute selection", In D.S. Huang *et al.*, (Eds.): ICIC **2009**, Lecture Notes in Artificial Intelligence, Springer-Verlag, vol. 5755, (**2009a**), pp. 91–100.
[3] T. Herawan and M. M. Deris, "Rough set theory for topological space in information systems", In the Proceeding of International Conference of AMS 2009, IEEE Press, (**2009b**), pp. 107–112.
[4] T. Herawan and M. M. Deris, "Rough set theory for selecting clustering attribute", In the Proceeding of International Conference of PCO 2009, American Institute of Physic, vol. 1159, (**2009c**), pp. 331–338.
[5] T. Herawan and M. M. Deris, "A construction of nested rough set approximation in information systems using dependency of attributes", In Proceeding of International Conference of PCO 2009, American Institute of Physic, vol. 1159, (**2009d**), pp. 324–331.
[6] T. Herawan, I. T. R. Yanto and M. M. Deris, "Rough set approach for categorical data clustering", In D. Ślęzak *et al.*, (Eds.): DTA 2009, Communication of Computer and Information Sciences, Springer-Verlag, vol. 64, (**2009**), pp. 188–195.
[7] T. Herawan, R. Ghazali, I. T. R. Yanto and M. M. Deris, "Rough set approach for categorical data clustering", International Journal of Database Theory and Application, vol. 3, no. 1, (**2010a**), pp. 33–52.
[8] T. Herawan, I. T. R. Yanto and M. M. Deris, "ROSMAN: ROugh Set approach for clustering supplier chain MANagement", Int'l Journal of Biomedical and Human Sciences, vol. 16, no. 2, (**2010b**), pp. 105–114.
[9] T. Herawan, M. M. Deris and J. H. Abawajy, "A rough set approach for selecting clustering attribute", Knowledge Based Systems, vol. 23, no. 3, (**2010c**), pp. 220–231.
[10] T. Herawan, I. T. R. Yanto and M. M. Deris, "A construction of hierarchical rough set approximations in information systems using dependency of attributes", In N.T. Nguyen *et al.*, (Eds.): Advances in Intelligent Information and Database Systems, Studies in Computational Intelligence, Springer-Verlag Berlin Heidelberg, vol. 283, (**2010d**), pp. 3–15.
[11] X. Huang, A. An, N. Cercone and G. Promhouse, "Discovery of interesting association rules from live link web log data", In the Proc. of IEEE Int'l Conference on Data Mining (ICDM'02), (**2002**), pp. 763-766.
[12] N. Labroche, N. Monmarche and G. Venturini, "Web sessions clustering with artificial ants colonies", In Poster Proceedings of the Twelfth International World Wide Web Conference (WWW'03), (**2003**).
[13] T. Li, Q. Yang and K. Wang, "Classification pruning for web-request prediction", In Poster Proceedings of the Tenth International World Wide Web Conference (WWW'01), (**2001**), pp. 1-2.
[14] C. Ling, B. S. Sourav and W. Nejdl, "COWES: Web user clustering based on evolutionary web sessions", Data and Knowledge Engineering, vol. 68, (**2009**), pp. 867-885.
[15] B. Mobasher, R. Cooley and J. Srivastava, "Creating adaptive web sites through usage-based clustering of URLs", In the Proc. Workshop on Knowledge and Data Eng'g Exchange (KDEX '99), (**1999**), pp. 19-25.
[16] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, "Effective personalization based on association rule discovery from web usage data", In Proceedings of the 3rd international workshop on Web Information and Data Management (WIDM'01), (**2001**), pp. 9-15.
[17] H. C. Morteza, A. Hassan and H. C. Mostafa, "Improving density based methods for hierarchical clustering of web pages", Data and Knowledge Engineering, vol. 67, (**2008**), pp. 30-50.
[18] Z. Pawlak, "Rough sets", Int'l Journal of Computer and Information Science, vol. 11, (**1982**), pp. 341-356.
[19] Z. Pawlak, "Rough sets: A theoretical aspect of reasoning about data", Kluwer Academic Publisher, **(1991)**.
[20] Z. Pawlak and A. Skowron, "Rudiments of rough sets", Inform. Sciences, vol. 177, no. 1, **(2007)**, pp. 3-27.
[21] N. Senan, R. Ibrahim, N. M. Nawi, I. T. R. Yanto and T. Herawan, "Rough Set Approach for Attributes Selection of Traditional Malay Musical Instruments Sounds Classification", In T.H. Kim *et al.*, (Eds.): UCMA 2011, Communication of Computer and Information Sciences, Springer-Verlag, vol. 151, (**2011a**), pp. 509–525.
[22] N. Senan, R. Ibrahim, N. M. Nawi, I. T. R. Yanto and T. Herawan, "Rough Set Theory for Feature Ranking of Traditional Malay Musical Instruments Sounds Dataset", In J.M. Zain *et al.*, (Eds.): ICSECS 2011, Comm. of Computer and Information Sciences, Springer-Verlag, vol. 188, no. II, (**2011b**), pp. 516–529.
[23] J. Srivastava, R. Cooley, M. Deshpande and P. -N. Tan, "Web usage mining: discovery and applications of usage patterns from web data", SIGKDD Explorations, vol. 1, no. 2, (**2000**), pp. 12-23.
[24] Q. Yang, H. H. Zhang and T. Li, "Mining web logs for prediction models in WWW caching and pre fetching", In Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01), (**2001**), pp. 473-478.
[25] I. T. R. Yanto, T. Herawan and M. M. Deris, "A framework on rough clustering for web transactions", In N.T. Nguyen *et al.*, (Eds.): Advances in Intelligent Information and Database Systems, Studies in Computational Intelligence, Springer-Verlag, vol. 283, (**2010a**), pp. 265–277.

[26] I. T. R. Yanto, T. Herawan and M. M. Deris, "RoCeT: rough set approach for clustering web transactions", International Journal of Biomedical and Human Sciences, vol. 16, no. 2, (**2010b**), pp. 135–145.
[27] I. T. R. Yanto, T. Herawan and M. M. Deris, "Data clustering using Variable Precision Rough Set", Intelligent Data Analysis, vol. 15, no. 4, (**2011**), pp. 465–482.
[28] I. T. R. Yanto, P. Vitasari, T. Herawan and M. M. Deris, "Applying Variable Precision Rough Set Model for Clustering Student Suffering Study's Anxiety", Expert System with Applications, vol. 39, no. 1, (**2012**), pp. 452–459.

Authors

Tutut Herawan

He received a B.Ed degree in year 2002 and M.Sc degree in year 2006 degree in Mathematics from Universitas Ahmad Dahlan and Universitas Gadjah Mada Yogyakarta Indonesia, respectively. He obtained a PhD in Theoretical Data Mining from Universiti Tun Hussein Onn Malaysia in year 2010. Currently, he is a lecturer with Department of Mathematics Education, Universitas Ahmad Dahlan, Indonesia**.** He currently supervises four PhD and had successfully co-supervised two PhD students and published more than 120 papers in various international journals and conference proceedings. He has appointed as an editorial board member for IJDTA, TELKOMNIKA, IJNCAA, IJDCA and IJDIWC. He is also been appointed as a reviewer of several international journals such as Knowledge-Based Systems, Information Sciences, European Journal of Operational Research, Applied Mathematics Letters, and guest editor for several special issues of international journals. He has served as a program committee member and co-organizer for numerous international conferences/workshops including Soft Computing and Data Engineering (SCDE 2010-2011 at Korea, SCDE 2012 at Brazil), ADMTA 2012 Vietnam, DTA 2011-2012 at Korea, DICTAP 2012 at Thailand, ICDIPC 2012 at Lithuania, DEIS 2012 at Czech Republic, NDT 2012 at Bahrain, ICoCSIM 2012 at Indonesia, ICSDE'2013 at Malaysia, ICSECS 2013 at Malaysia, SCKDD 2013 at Vietnam and many more. His research area includes Knowledge Discovery in Databases, Educational Data Mining, Decision Support in Information System, Rough and Soft Set theory.

Wan Maseri Wan Mohd

He received a B.Sc. in Computer Science from University of Maimi Florida USA, 1985, a M.Sc. in Computer Science from University of Maimi Florida USA, 1986 and a Ph.D. in Management, University Technology Malaysia. Currently she is an associate professor at Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang. She is also the director of ICT Business Center, Universiti Malaysia Pahang. She has published more than 30 research papers in journals and conferences. Her research interest includes knowledge management, information retrieval and data mining.

